**Sequence and dependencies**

1. **Prepare datasets from national databases**

Books\_BEdata\_NPR\_prep.R

Books\_BEdata\_prep.R

Books\_HRdata\_addLanguage.R

Books\_HRdata\_prep.R

Books\_SIdata\_extractISBNsfromXML.R

Books\_SIdata\_prep.R

Books\_NOdata\_prep.R

Input:

Raw data from national databases

Output:

1. A list of ISBNs
2. A list of unique ISBN and local ID pairs
3. A dataset of unique local IDs with metadata
4. **Combine ISBNs for OCLC**

Books\_ISBNs\_Combined.R

Input: 1a (for OCLC)

Output: A list of unique ISBNs (n=19182)

!Note: ISBNs are standardised to ISBN13 and validated

1. **Combine metadata and ID-ISBN lists, and export datasets with unique language and publisher values**

Books\_metadata\_Combined.R

Input: 1b, 1c

Output:

1. Combined ID-ISBN pairs (n=20465)
2. Combined metadata (n=19332)
3. A dataset of unique language values (n=405)
4. A dataset of unique publisher values (n=5342)
5. **Clean data from Worldcat**

Books\_Clean\_WorldCat-data.R

Input:

* the two files provided by OCLC
* combined metadata
* combined ID-ISBN pairs

Output:

1. cleaned worldcat d1 (ISBNs and oclns; n=117620)
2. cleaned worldcat d2 (oclns and ISBNs; n=103943)
3. a dataset for checking isbn-ocln link accuracy (contains ISBN and ocln links and metadata from worldcat and national dbs; n=125617)

!Note: ISBNs in worldcat d1 are standardised to ISBN13 and validated

1. **Add ISBNs from OCLC to ISBNs from national databases + identify local IDs with 6 or more ISBNs**

Books\_Add\_cleaned\_OCLC\_ISBNs.R

Input:

* Cleaned worldcat d1 (4c)
* Manually cleaned worldcat d1 (a version of 4c from August 2019)
* Cleaned worldcat d2 (4b)
* Combined metadata
* Combined ID-ISBN pairs

Output:

1. localID, source, and ISBN combinations (all ISBNs: from national DB and OCLC; n=32679)
2. records with 6 or more ISBNs (n=634).
3. **Identify potential bibliographic families**

Books\_Potential\_biblio-families.R

Input:

* Combined metadata
* localID, source, and ISBN combinations (5a)

Output: a dataset with local IDs, ISBNs and metadata for manual validation of ISBN overlap (n=4868)

1. **Scrape Goodreads**

Books\_Goodreads\_addMetadata.R

Input: localID, source, and ISBN combinations (5a)

Output: a sequence of datasets containing metadata retrieved from Goodreads (by ISBN)

Note: This script is not run again as it takes enormous amount of time (~week) and also the Goodreads records continuously are being dropped (some of records that were retrieved in August 2019 were no longer there in December 2019 or now, in June 2020).

1. **Clean Goodreads data**

Books\_Goodreads\_Combined.R

Input:

-datasets with metadata retrieved from Goodreads,

- localID, source, and ISBN combinations (5a)

Output:

-Goodreads metadata combined with local IDs and Source; includes also information on whether the record has been valid when checked manually (only for multiISBN records); n=32679.

1. **Clean metadata**

Books\_Clean\_metadata.R

Input:

-Combined metadata

-Manually cleaned publisher file

-Manually cleaned language file

-Manually validated files with records that contains 6 or more ISBNs

-Dataset with information on COBISS authors

Output:

-LocalID, Source, Cleaned language, Cleaned publisher, COBISS author info, multiISBN info.

1. **Assign work identifiers to localIDs**

Books\_Assign\_workIDs-to-localIDs.R

Input:

-Combined metadata

-Manually cleaned bibliographic family file

-File with potential bibliographic families

Output: localID, source, workID (n=19332)

Note: The new file of bibliographic families contains much more records than the earlier file. I don’t know the reason why there are more records now, but it seems to be correct.

Note: a couple of CRISTIN records are not exported correctly; before the manual exploration in MS Excel, I edited the file in Notepad++.

Note: n records now identified as works contain 6 or more ISBNs but have not been checked manually (whether its multi ISBN or Misc)

1. **Prepare datasets for analysis**

Books\_Datasets-for-analysis.R

Input:

-All ISBNs with local IDs (5a)

-Combined metadata (3b)

-Cleaned metadata (9)

-Cleaned WorldCat data (4b)

-Cleaned Goodreads data (8)

-WorkIDs (10)

Output:

1. ISBN as a unit of analysis
2. Local ID as a unit of analysis
3. Work ID as a unit of analysis
4. Group2
5. Group3
6. Sample of work with many ISBNs
7. **Analyse the data**

Books\_Analysis.R

Input: 11a, 11b, 11c